NEAT THINGS

CHRISTIAN PERLE

Unsure whether the HTML on your own web pages is right? Has the HTML export from word processing or the HTML editor produced gobbledegook? With *tidy* you can clear it up.

HTML: "*H*yper*T*ext *M*arkup *L*anguage", the mark up language originally developed at CERN for sites on the World Wide Web. By using so-called *Tags*, specified sections of text are marked up as headings, lists, tables and suchlike.

WYSIWYG: "*W*hat *Y*ou *S*ee *I*s *W*hat *Y*ou *G*et", a concept in popular use in the office domain, in order to see inputs immediately in the formatted representation. Since in the case of HTML the exact appearance of headings and other page elements is not defined, it is not exactly compatible with the WYSIWYG concept.

Out of the box

There are thousands of tools and utilities for Linux. "Out of the box" takes the pick of the bunch and suggests a little program each month which we feel is either absolutely indispensable or unduly ignored.

Anyone putting their own websites onto the Net creates these in one of two ways. Whether with a simple text editor or with an **HTML** editor, which works according to the **WYSIWYG** principle, in either case errors can arise. With *tidy*, there is now a tool available to help you to create "clean" and standardised HTML documents.

Listing 1: HTML with errors

```
<title>Sloppy page
<h1>A page full of errors</h2>
... and that means food for <i>tidy</i>!
<P>
On this page there are
<LI> wrong and missing tags,
<li>an incomplete list
<li>and <i>wrongly <B>nested</i></B> tags.
```

Listing 2: Tidy has cleared up

```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 3.2//EN">
<HTML>
<HEAD>
<META name="generator" content="HTML Tidy, see www.w3.org">
<TITLE>Sloppy page</TITLE>
</HEAD>
<BODY>
<H1>A page full of errors</H1>

...and that means food for <I>tidy</I>!

<P>On this page there are</P>

<UL class="noindent">
<LI>wrong and missing tags,</LI>

<LI>an incomplete list</LI>

<LI>and <I>wrongly <B>nested</B></I> tags.</LI>
</UL>
</BODY>
</HTML>
```

Perfect nonsense

One often reads, on WWW pages, the phrase "optimised for Netscape Navigator" or "optimised for Internet Explorer". But in point of fact, this alleged optimisation really means that anyone using a different browser can look forward to incorrectly or incompletely displayed page contents. It's better to use valid standards supported by the majority of browsers.

It's not without good reason that *tidy* keep very closely to HTML standards. It is being developed by Dave Raggett in the frame of the WWW Consortium (W3C), which is working out precisely these standards. But before *tidy* clears up for you, you have to install it.

From a reliable source

To do this, get the source archive from the *tidy* home page (*http://www.w3.org/People/Raggett/tidy/*) and install it with the following commands:

```
tar xzf tidy4aug00.tgz
cd tidy4aug00
make
su
(enter root password)
make install ; exit
```

If this procedure has gone smoothly, you can subject tidy to a first operational test.

Make it better

Using a text editor, create the file *sloppy.html (Listing 1)*. You can now let *tidy* loose on this HTML catastrophe and divert the improved result into the file *better.html*:

```
tidy -upper sloppy.html > better.html
```

The result can be seen in Listing 2. The program also issues some error messages in addition to the corrected output, so you can understand exactly what it was that *tidy* did not like and what has been added or replaced. The option *-upper* ensures

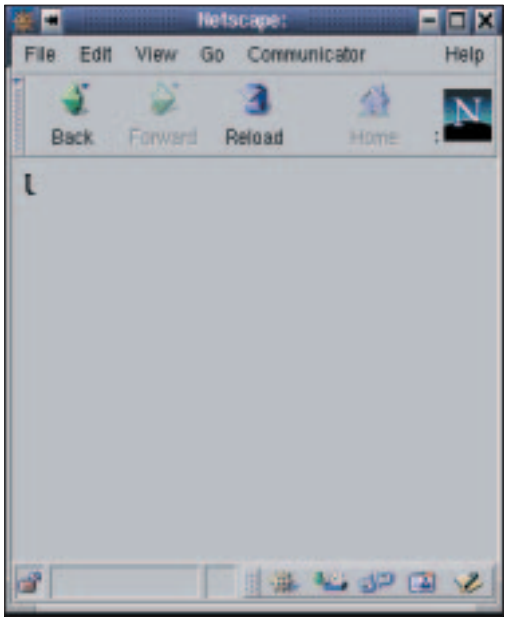

Figure 1: Netscape is confused

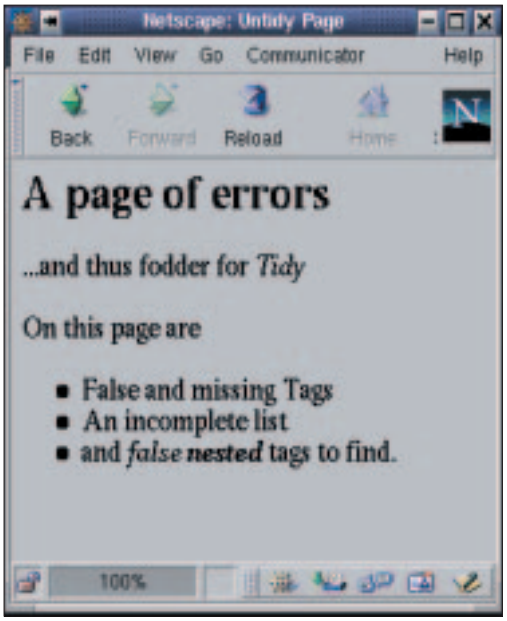

Figure 2: tidy works wonders

that all tags are written out at the same size. The corrections include, among others:

- Completing the page with HTML, HEAD and BODY tags,
- Closing the TITLE tag (a significant difference, as can be seen in Figures 1 and 2),
- Correcting the H1 heading erroneously closed with </H2>,
- Enclosing the list tags with a UL tag,
- Closing each individual list tag with </LI>,
- Swapping the closing tags </I> and </B> for correct nesting and
- Replacing the umlauts with **HTML entities**.

The last point is necessary because of the DOCTYPE comment added by *tidy*. If *DE* crops up here instead of *EN*, then **ISO-Latin-1** coded umlauts are permissible. In order to leave such umlauts unchanged in the *tidy* run, use the option *-raw*.

Beautiful HTML

In order to format computer-created HTML to make it easily legible for manual post-editing, *tidy* has the *-indent* option, which indents the respective HTML elements according to their depth of nesting. In Listings 3 and 4, the effect of this option is demonstrated.

If you allow table elements to indent in the same way, their representation can easily be changed due to errors by some browsers. So it's better if you control the result.

Even the HTML derivative interspersed with proprietary extensions which is created by MS Office 2000 can be put into a clean form which is also more suitable for the WWW by *tidy* (Option *—word-2000 yes*). In the test a document 90 KB long was reduced to one tenth(!), without any sacrifice of data content.

Fine-tuning

If you need certain command line options of *tidy* again and again, but are tired of constantly typing them in, entries in a configuration file would be a good idea. It's up to you whether you want to use a system-wide configuration or one which is linked to your user account. To tell *tidy* where the configuration file is, set the **environment variable** *HTML_TIDY* to the corresponding file name, such as */etc/htmltidy.conf*. To do this, add to your *.bashrc* the line *export HTML_TIDY=/etc/htmltidy.conf*.

Now enter your standard options into the configuration file. Listing 5 shows one example. A wide-ranging overview of the options can be found in the file *Overview.html* in the *tidy* ource archive.

With *wrap: 72*, lines in the HTML document are broken after 72 characters, *indent: auto* gives automatic indenting with the exception of tables, *char-encoding: latin1* selects the character set coding and *uppercase-tags: yes* does the same as the option *-upper*. I can strongly recommend the home page of the program for additional functions and options of *tidy*.

Listing 3: Structure not recognised...

```
<ol><li>A<ul><li>a<li>fat</ul>
<li>oily <li>
fish <li>went angling</ol>
```

Listing 4: ... but now it is

```
<ol>
 <li>
 A
 <ul>
  <li>fat</li>
  <li>oily</li>
 </ul>
</li>
<li>fish</li>
<li>went</li>
<li>angling</li>
</ol>
```

Listing 5: Example of htmltidy.conf

```
wrap: 72
indent: auto
char-encoding: latin1
uppercase-tags: yes
```

HTML entities: A substitute notation for characters not included in the 7-bit ASCII character set or characters which have a special meaning in HTML. The entity for the umlaut is *ü*, that for the copyright symbol *©* or that for the less-than symbol *<*.

ISO-Latin-1: A standard for the coding of country-specific and special characters as extension of the ASCII character set. The latter prescribes only the codings from 0 to 127 and thus leaves out e.g. umlauts.

Environment variables: These variables automatically pass certain system settings to processes, for example the search path for programs *PATH*, the localisation settings *LC_LANG* and *LC_CTYPE* or again, the name of a configuration file.